160-WP-003-001

A Review of EOSDIS QA Metadata: Supprt for EOS Quality Assessment (QA) and QA Metadate Update Tools

April 1999

Note: This paper was published in April 1999 for the IEEE Metadata Conference.

White Paper--Not intended for formal review or government approval.

A Review of EOSDIS QA Metadata: Support for EOS Quality Assessment (QA) and QA Metadata Update Tools

[1] **Sushma Singhal (Raytheon ITSSCorp.) and Bob Lutz (Raytheon ITSS Corp.)**

ABSTRACT

Quality Assessment (QA) is defined within NASA Earth Observing System (EOS) as the process that identifies and flags data products that obviously and significantly do not conform to the expected accuracies for the particular product type. Operationally, QA involves quality control measures that can be applied to the data products, before release to the general science community. This is a challenging task within EOS, due to the large volume and the complexity of data produced and the near real time mode between data production and distribution. The EOS Data and Information System (EOSDIS) provides the computing and network facilities, to support the generation and archival of data products from EOS satellites, and distribution of data productions to user community. To accommodate the high volume data production and simultaneously provide end users with some indication of scientific and operational quality of data, the ECS has incorporated a high degree of automation within its design. EOS QA methodology integrates: (I) the automated detection of certain types of suspect data during product generation, (ii) the capability of EOSDIS to alert the instrument team scientists and processing facility personnel to suspect data, (iii) the extraction of this data from the archives for QA purposes and the subsequent storage of QA results within EOSDIS, and (iv) the organization, archival and display of all of these QA results in a user-friendly format for the user community.

This paper discusses the ECS metadata attributes designed to support QA functions, and tools and functionality that have been developed by ECS for the science teams and the DAACs to perform their manual QA analyses and enter their QA results into the QA Metadata.

Keywords: Collection, DAAC, Distributed Active Archive Centers, Granule, PGE, Product Generation Executives, QA analyses, Quality Assessment, Metadata QA

1. INTRODUCTION

The Earth Observing System (EOS) [1] consists of a space flight component, a science component and a data system. The space flight component is a coordinated series of polar-orbiting and low-inclination satellites for long-term global observations of the land, atmosphere,

[1] Authors contact information

Sushma Singhal can be contacted at (301) 883-4097; ssinghal@eos.hitc.com

Bob Lutz can be contacted at (301) 614-5037; rlutz@ltpmail.gsfc.nasa.gov

and oceans. The first operational version of the EOSDIS for the NASA's Earth Science Enterprise will be released during 1999. This will be the world's largest civilian data management system for remotely sensed data. The AM-1 spacecraft will be the first comprehensive satellite of EOS scheduled to be launched in summer of 1999. The spacecraft consists of five instruments: the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), the Clouds and Earth’s Radiant Energy System (CERES), the Multi-Angle Imaging SpectroRadiometer (MISR), the Moderate-Resolution Imaging Spectrometer (MODIS), and the Measurements of Pollution in the Troposphere (MOPITT). There is an associated instrument team (IT) for each instrument developing the science algorithms and processing software. MODIS, producing more data than the four other instruments combined, has their instrument team split into three disciplines: atmospheres, oceans and land. The instrument teams and their programming staff use one or more science computing facilities (SCFs) to develop and test the science algorithm software and to support IT quality control analyses.

The Earth Observing System Data and Information System (EOSDIS) [2] is a NASA-sponsored open, distributed information system that will manage the data and information from a variety of pre-EOS and EOS-era Earth observation satellites, as well as data from related Earth science field measurement programs and other data essential for the interpretation of these measurements. EOSDIS will provide end-to-end services from EOS instrument data collection to science data processing to full access to EOS and other Earth science data holdings. The EOSDIS Core System (ECS) is the infrastructure of EOSDIS that provides the computing and network facilities to support the generation, archival and distribution of geophysical data products generated from the data sensed by the EOS satellites.

The EOSDIS has been under development to support the AM-1 spacecraft and future EOS missions. EOS AM-1data will be generated and archived at four operational Distributed Active Archive Centers (DAACs): Goddard Space Flight Center (GSFC) —MODIS; Langley Research Center (LaRC) —CERES, MISR, MOPITT; EROS Data Center (EDC) —MODIS, ASTER; and the National Snow and Ice Data Center (NSIDC) —MODIS. In addition, EOSDIS will also archive geophysical parameters generated at processing facilities outside of the ECS environment (e.g., at Principal Investigator (PI) processing facilities)

One of the mandates of the EOS data policy is that all data be available to the science community in a timely manner. Quality Assessment (QA) of EOS data therefore is a critical component of the processing system because suspect and bad data must be flagged before data products are released to user community. The ECS provides the instrument team scientists the computing architecture needed to quality assess their data (e.g., the Client tool — used for EOSDIS data search and order).

The EOS QA process identifies and flags data products that obviously and significantly do not conform to the expected accuracies for the particular product type. This is a challenging task within the EOS due to the large volume of data produced (one terabyte per day), the near-real-time mode between data production and distribution, and the numerous error sources that may effect data quality. To accommodate the high volume data production and simultaneously provide end users with some indication of scientific and operational quality of data, the ECS has incorporated a high degree of automation within its design. To support EOS Quality Assessment (QA) efforts, ECS integrates: (a) the capability to alert the science teams and processing facility personnel to suspect data through automated flagging within the algorithm software, (b) the retrieval of this data from the archives for QA purposes and the subsequent insertion of manual QA results within EOSDIS, and (c) the organization, archival and display of all of these QA results in a user-friendly format for the user community. QA Metadata, being part of the ECS Metadata model, has been developed to support the above objectives.

Within the Product Generation Executives[2] (PGE), automated QA may be performed, to ensure at least the minimum needed quality control of all data. These automated QA results are stored within the metadata, with values determined by the quality assessment criteria defined by the science teams. Some limited QA is performed manually on subsets of the data products by staff at the processing facilities and the science team facilities. These QA results are also captured and stored in the metadata.

The following sections provide an overview of the ECS metadata attributes designed to support QA functions and discuss the tools and functionality that have been developed by ECS for the science teams and the processing facilities to perform their manual QA analyses and enter their QA results into the QA metadata.

2. ECS QA METADATA

The ECS Metadata model, consisting of more than 280 "core" Metadata attributes[3], includes several QA information attributes to reflect the quality assessments of the data products. The QA process is performed primarily at the granule or smaller level, where a granule is defined as the smallest entity of a data set that is tracked and managed by the system. QA information attributes exist for both granule[4] and collection[5] (e.g., data set) metadata. Collection-level QA information attributes includes pointers including URLs to QA related documents (User Comment Document) and review material. Granule-level QA information attributes includes multiple QA Flags (e.g., Automatic quality flag), QA statistics (e.g., Percent missing data), processing QA description, pointers to browse products, and QA granules supplied by data producers. In addition to the above core attributes, the ECS metadata model provides product specific extensions which are used by the science teams to provide QA information unique to a particular product. Product-specific attributes are used to describe specific characteristics of the instrument at the time of sensing, or information that applies only to a certain discipline, or information that is agreed as important for a smaller segment of the science community rather than across-the-board, can be provided in addition to the core.

[2] A Product Generation Executive (PGE) is defined as the smallest schedulable unit of science software in the ECS production system. The science teams provide the PGE software, which is comprised of one or more executables.

[3] The ECS term ‘core’ metadata is defined as the set of common attributes, which must exist for all ECS standard data products where applicable.

[4] A granule is defined by ECS as the smallest entity of a data set that is described, tracked, inventoried, and managed by the system. A granule consists of one or more physical files.

[5] A collection is defined by ECS as a logical grouping of granules chosen by data providers for publishing in ECS as a collection. A granule may end up “belonging” to several collections.

Granule level QA metadata consists of core (all products) and non-core (product specific) metadata [3] attributes as listed in Table 2.1.

Table 2.1-1: Granule Level QA Metadata

Type of Metadata	**Group**	**QA Metadata**	DataType

Core QA Metadata (common to all products)	QAFlags	AutomaticQualityFlag
		OperationalQualityFlag
		ScienceQualityFlag
			Varchar(64)
			Varchar(25)
			Varchar(25)

	QAFlagsExplanation	AutomaticQualityFlagExplanation OperationalQualityExplanation ScienceQualityFlagExplanation	Varchar(255) Varchar(255) Varchar(255)
	QAStats	QAPercentMissingData QAPercentOutofBoundsData QAPercentInterpolatedData QAPercentCloudCover	Float Float Float Float
Non-core (Product specific attributes [PSAs])		QA PSAs —defined by the ITs	

2.1 Core QA Metadata

Core metadata, being common to all EOS products, allows the user to utilize global search criteria for browsing and searching the EOSDIS database. It has been arranged in a simple and concise format to allow the greatest utilization by the science community. There are two components of QA core metadata: QAFlags and QAStats.

2.1.1 QAFlags

A set of three general QA flags is used to indicate the overall quality assessment level of the granule. Text comment fields (QAFlagExplanation) are available to supplement these flags.

a. **AutomaticQualityFlag —**This flag is used for automatic quality assessment of data products during products generation and is set by the algorithm processing software within the PGE. The valid values are Passed, Failed, and Suspect. Criteria for setting this flag (e.g., What constitutes "Passed"?) are determined by the ITs. There is no default valid for this flag and it must be set in the PGE.

b. **OperationalQualityFlag—**This flag is used for manual QA and may be set by processing facility personnel (DAAC or PI) to indicate the results of non-science QA (i.e., data are not corrupted in the transfer, archival and retrieval process). The valid values are Passed, Failed, Suspect, Being Investigated, Not Investigated, Inferred Passed, and Inferred Failed. Not Investigated is the default value assigned by the system, if non-science QA is not performed.

c. **ScienceQualityFlag—**This flag used for manual QA and is set by the IT scientists or their designees (e.g., personnel at the processing facility) indicating the results of science QA. The valid values are the same as the OperationalQualityFlag, with the addition of

"Validated"—the granule has been validated by an expert (e.g., the granule has been compared to in-situ data). The default value is Not Investigated.

2.1.2 QAStats

A set of generic numerically based flags that are associated with each granule. These attributes are:

a. QAPercentMissingData

b. QAPercentOutofBoundsData

c. QAPercentInterpolatedData

d. QAPercentCloudCover

These parameters are generated within the PGEs, with values that range from 0 to 100 or a default null value. The instrument team scientists writing the algorithm software determine the criteria. Some teams are opting not to populate all of these parameters where they believe the parameters are not meaningful for their specific products (e.g., the MISR team does not populate the QAPercentOutofBoundsData or QAPercentCloudCover parameters). In addition, some of these flags may not be informative for all levels of data (e.g., all Level 3 data are interpolated data).

2.2 Product Specific QA Metadata

To indicate individual product QA information, specific granule-level QA parameters are established by the ITs. These parameters are assigned values within the PGEs and are known as QA product specific attributes (QA PSAs). Being part of the non-core metadata, these parameters may also be utilized, in addition to the core metadata, by the user for data searches within EOSDIS. Many of the QA PSAs defined for the granule may be summary statistics of the sub-granule (e.g., pixel level) QA parameters. For example, for the MODIS Atmospheric Aerosol product, two of the defined QA PSAs are "percent success rate of retrieval—land" and "percent success rate of retrieval—ocean."

3. EOS Metadata Quality Assessment

Though the specifics may vary for each IT's QA scenario [4], there are two general methodology to perform operational QA: automated and manual operational QA

3.1 Automated Operational QA

The automated QA is performed by the algorithm software (the PGEs), which generates the products (PGE QA analysis). The data products are produced at a processing facility (DAAC) or PI from science algorithms supplied by the instrument science teams. Numerous QA parameters (operational and product-related) are generated by these algorithms. These generated QA parameters may be at the granule or sub-granule level, and possibly summarized or subsetted. These QA parameters are then sorted and subdivided among the product metadata, the data product and any external QA products. From criteria specified by the instrument teams, the core metadata field—the AutomaticQualityFlag (flag and text)—is set within the PGEs. In addition, some teams (e.g., ASTER) are making extensive use of alerts or alarms in their processing software to warn them of anomalous conditions that occur during production.

3.2 Manual Operational QA

Some limited manual QA on a subset of data products may be performed by personnel or software at the DAAC or PI processing facility (Processing Facility QA Analysis) and/or by IT scientists or their designees at the Science Computing Facility (SCF QA Analysis)

3.2.1 Processing Facility QA Analysis

The processing facilities are responsible for monitoring non-science QA aspects of data production. They are to check the integrity of the data at the file level, to ensure that the data are not corrupted in the transfer, archival, or retrieval processes. This analysis may include checking that the file can be opened and that the file size is correct. In addition, some processing facilities may perform limited science QA functions, in agreement with their ITs. This may involve monitoring summary QA statistics and alerts generated from the PGE QA analysis or visually displaying data to detect gross problems. The results from non-science QA analyses performed at the processing facilities are summarized in the core metadata OperationalQualityFlag, and text field by authorized processing facility staff.

3.2.2 SCF QA Analysis

The ITs ultimately are responsible for the science QA of their data products. Each instrument team has developed a different strategy and set of procedures to accomplish this objective. There are two general types of QA analyses performed by instrument team scientists: 1) those of an investigative nature and principally analyzing suspect data and 2) those of a routine nature, involving regular screening of the data product. Many teams are estimating that they can routinely examine 10% of the daily averaged data production. It is expected that, during the first year, a greater emphasis will be placed on analyzing suspect data. Maturity in the understanding of the behavior of the instruments, and revised science algorithms, should see a gradual change from investigative QA to routine QA screening in later years. Scientists at the SCF may examine a subset, or the entire data product stream for instruments with low data rates. For most AM-1 instrument teams it is impractical to transfer the full set of data products from the DAAC to the SCF because of prohibitively large network requirements. Therefore, over a given time period, most teams intend to order only statistical samples and samples of those data with quality problems indicated by their QA metadata (core and PSA) The results of science QA analyses performed at the SCF are summarized in the core metadata ScienceQualityFlag and text field by the instrument team scientists.

4. QA Metadata Update Tool

Initially the QA Monitor (not described in this paper) was developed by ECS, to support the insertion of QA metadata updates into the DAAC Science Data Server. When using the QA Monitor, metadata insertion updates must be done manually one granule at a time by a person at the DAAC. Therefore, the use of the QA Monitor was deemed not to be practical or adequate for the expected hundreds of QA metadata updates generated by the science teams (and possibly at the processing facilities) and received each day by the DAACs. Therefore a tool was developed by ECS to support batch updating of the QA Metadata - the QA Metadata Update Tool (QAMUT).

4.1 QA Metadata Update Tool

The QA Metadata Update Tool (QAMUT) allows an authorized user (e.g., an IT scientist) to perform batch updates of QA metadata, It enables both SCF and processing facility QA experts to modify values of their respective quality flags (i.e., ScienceQualityFlag and OperationalQualityFlag) on core, provided via a search Client interface, for multiple granules at a time in a batch mode. The Java Earth Science Tool (JEST) Client, was initially designed and planned to provide the data search and retrieval functionality and configured as the front end of QAMUT, but plans are now to use the enhanced V0 Client. The QAMUT provides a Web based Graphical User Interface. This is both for ease of use and to simplify the interface implementation. Access to the QAMUT is limited to persons authorized to update metadata for certain data types only. Based on the user's privileges in the user profile, users can update either the ScienceQualityFlag, the OperationalQualityFlag, or both

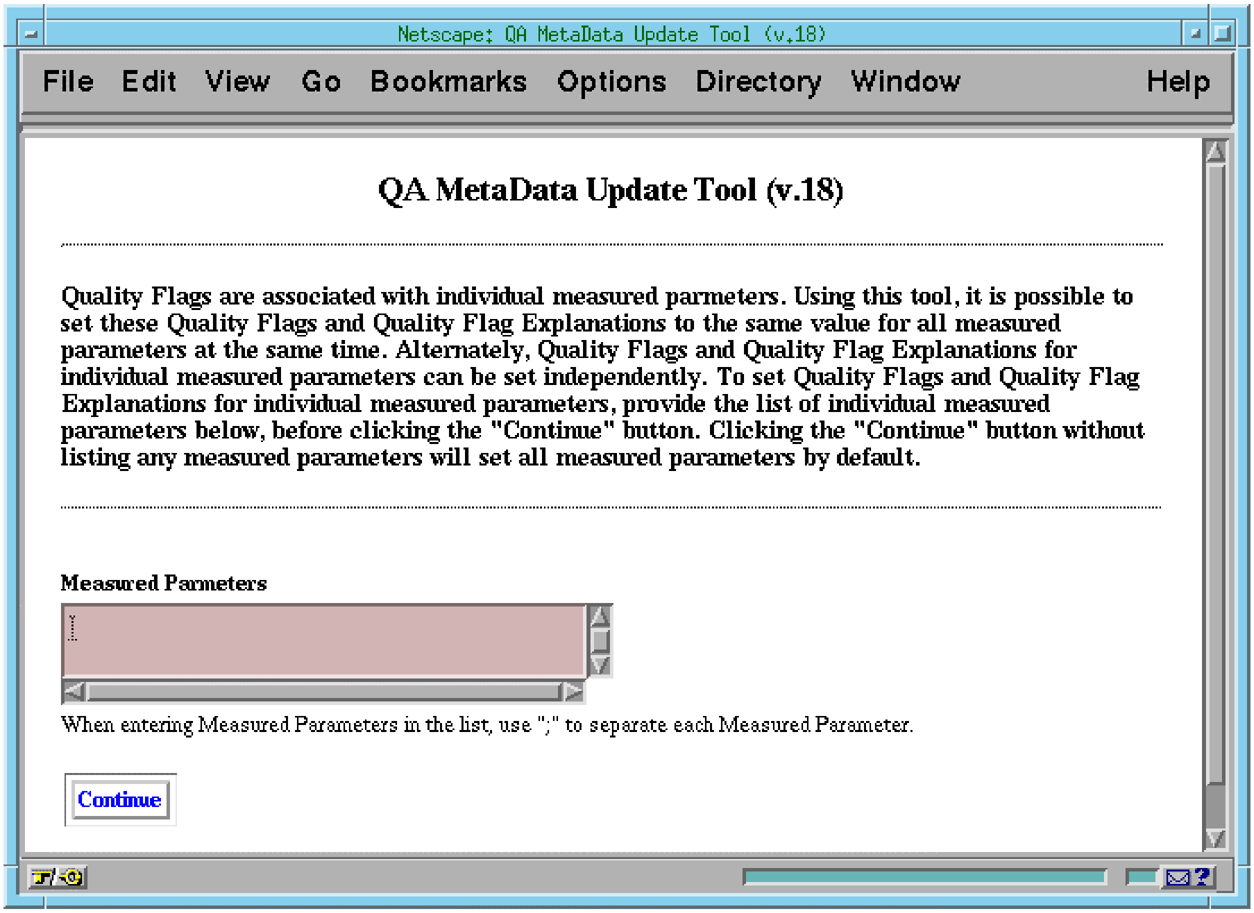
The tool consists of two major components: The front end linked to the Client (SCF Metadata Tool) and the back end, linked to the DAAC Data Server (DAAC Operator Tool).

4.1.1 SCF Metadata Tool

The Client is initially utilized to derive a set of granules for which the quality flags and/or text will be updated. Within the ECS B.0 Metadata model, a granule can contain multiple geophysical parameters.

Each geophysical parameter has the full complement of QAFlags and, if appropriate, QAStats. The SCF Metadata Tool displays the set of granules obtained from the query of the Client as a listing of records, where there may be multiple records for a granule corresponding to the number of parameters.

The information (QA Flags and text) is displayed sequentially on the screen for user to see and to alter interactively. The Figure 4.1-1 shows the initial screen layout for the MUT interface.



QA MetaData Update Tool (v.18)

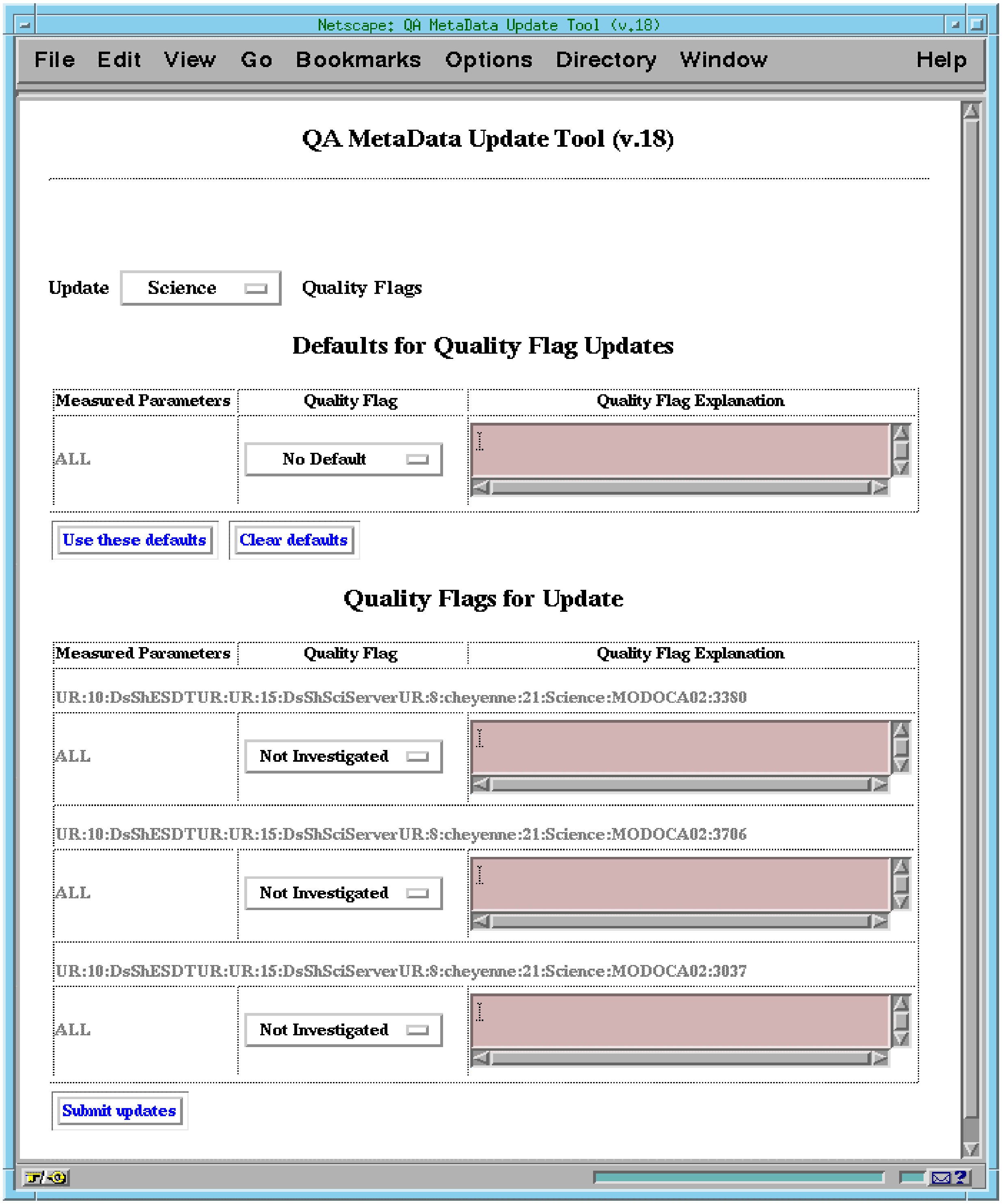
Quality Flags are associated with individual measured parmeters. Using this tool, it is possible to set these Quality Flags and Quality Flag Explanations to the same value for all measured parameters at the same time. Alternately, Quality Flags and Quality Flag Explanations for individual measured parameters can be set independently. To set Quality Flags and Quality Flag Explanations for individual measured parameters, provide the list of individual measured parameters below, before clicking the "Continue" button. Clicking the "Continue" button without listing any measured parameters will set all measured parameters by default.

Measured Parmeters

When entering Measured Parameters in the list, use ";" to separate each Measured Parameter.

Figure 4.1–1. The QAMUT User Screen Layout

On this screen the user enters a list of measured parameters, if desired. By not providing a list of measured parameters, the user by default will update all measured parameters for each granule. Clicking the continue button, brings up the screen to select the quality flag and quality flag explanation for the granules as shown in the Figure 4.1-2



QA MetaData Update Tool (v.18)

Update Science Quality Flags

Defaults for Quality Flag Updates

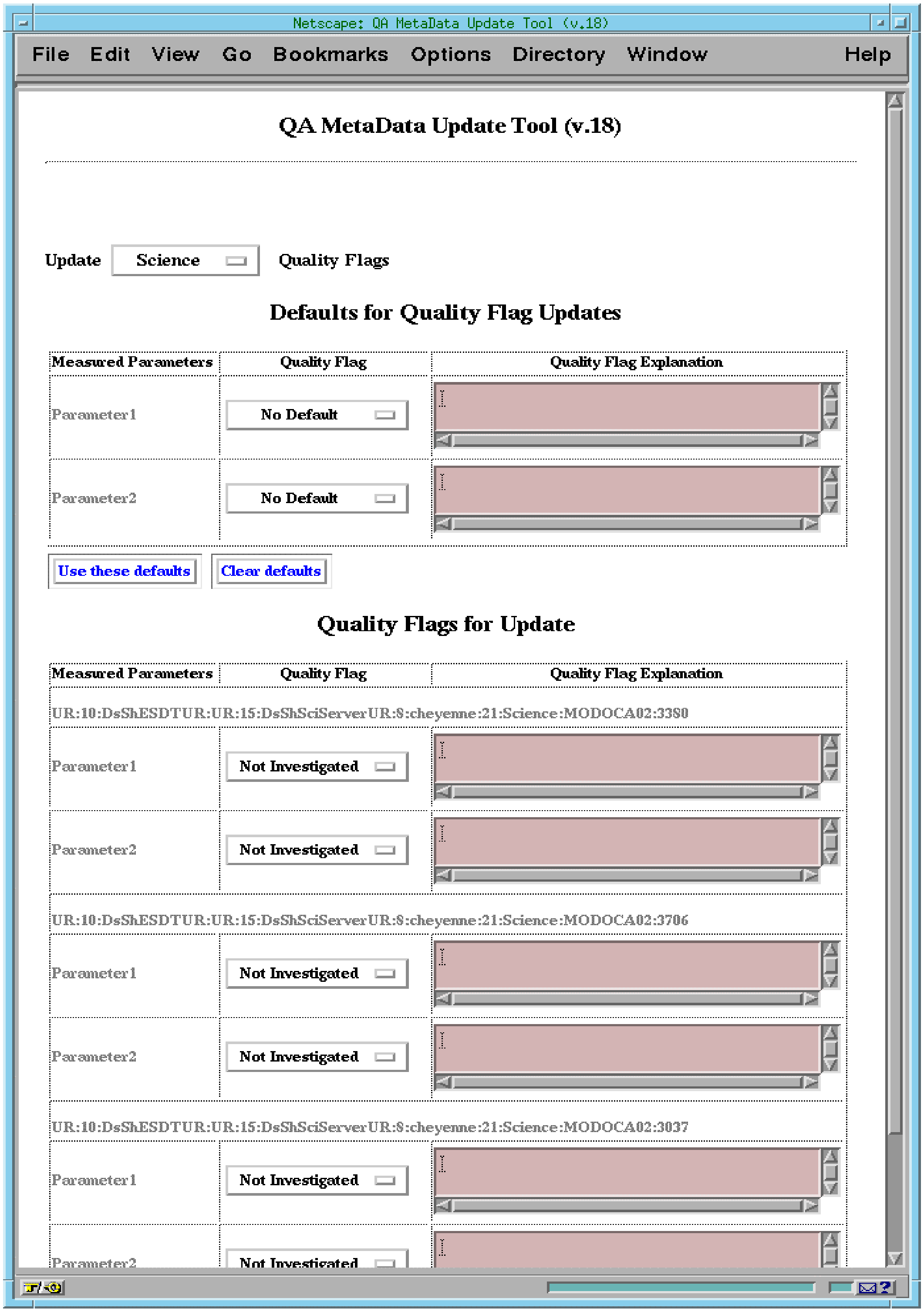
Measured Parameters	Quality Flag	Quality Flag Explanation
ALL	No Default	

Quality Flags for Update

Measured Parameters	Quality Flag	Quality Flag Explanation
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3380		
ALL	Not Investigated	
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3706		
ALL	Not Investigated	
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3037		
ALL	Not Investigated	

Figure 4.1-2. The QAMUT User Screen Layout

The QA metadata updates on all displayed data are performed by selecting a value from a pull-down menu. The tool allows updates to all records in the result set by substituting user-specified valid value for the existing ScienceQualityFlag values. The Figure 4.1-3 shows the modified values.



QA MetaData Update Tool (v.18)

Update: Science Quality Flags

Defaults for Quality Flag Updates

Measured Parameters	Quality Flag	Quality Flag Explanation
Parameter1	No Default	
Parameter2	No Default	

Quality Flags for Update

Measured Parameters	Quality Flag	Quality Flag Explanation
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3380		
Parameter1	Not Investigated	
Parameter2	Not Investigated	
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3706		
Parameter1	Not Investigated	
Parameter2	Not Investigated	
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3037		
Parameter1	Not Investigated	
Parameter2	Not Investigated	

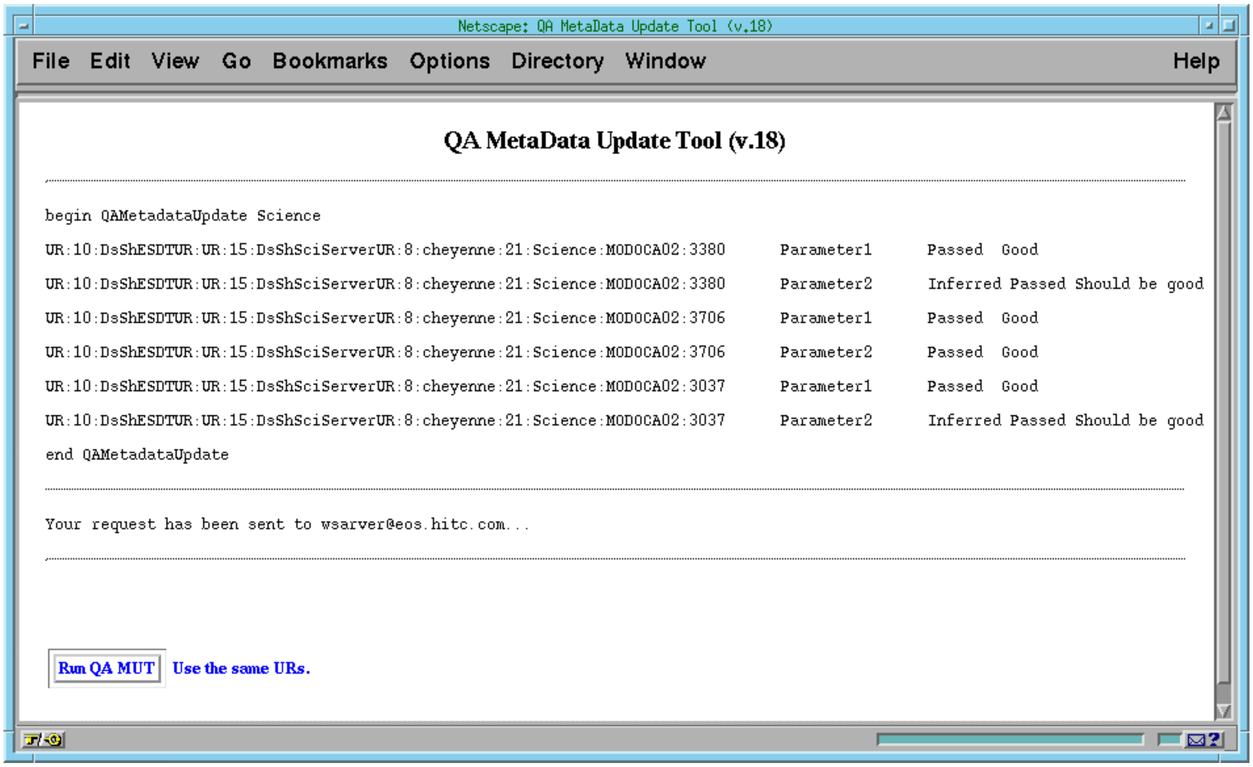
Figure 4.1-3. The modified values to QA Metadata

Presently, the domain of valid values for the ScienceQualityFlag include "Passed", "Failed", "Being Investigated", "Not Investigated", "Inferred Passed", "Inferred Failed". And "Validated". A desirable feature of the tool is that it allows users to modify that single value to exceptions for those granules for which the standard value is incorrect. A "Forward QA Updates" button is provided on the SCF QA Update interface. This button invokes the CGI e-mail program. The underlying code formats the message using the modified result set. The update history is maintained for posterity and later may be used for trend analysis, if required. Any correction to unintended updates will require a new QA Update request submission to the DAAC Operator.

4.1.2 DAAC Operator Tool

The main objective of the DAAC operator tool is to actually perform the metadata updates as received from the SCF Metadata Tool. Specifically, the DAAC operator tool will parse the batch metadata update list provided by the SCF tool and initiate the changes in Science Data Server. The DAAC operator tool includes the functionality to go through Data Server interfaces to trigger any metadata update events.

The Figure 4.1-4 shows the update message as received by the DAAC operations. At the initiation of the tool, the operator views the list of metadata updates vs. granule and MeasuredParameter.



QA MetaData Update Tool (v.18)

```
begin QAMetadataUpdate Science
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3380     Parameter1     Passed  Good
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3380     Parameter2     Inferred Passed Should be good
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3706     Parameter1     Passed  Good
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3706     Parameter2     Passed  Good
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3037     Parameter1     Passed  Good
UR:10:DsShESDTUR:UR:15:DsShSciServerUR:8:cheyenne:21:Science:MODOCA02:3037     Parameter2     Inferred Passed Should be good
end QAMetadataUpdate

Your request has been sent to wsarver@eos.hitc.com...
```

Figure 4.1-4. The Update Message

Similar to the SCF user interface, the DAAC operator interface requires a DAAC operator to have an authorized DAAC operator account.

The DAAC operator tool waits until after the requested QA metadata updates are committed in the Science Data Server databases and then sends notification back to the user indicating updates were made. This provides the user with timely feedback as well as an independent communication mechanism between the user and the DAAC.

The update history is maintained by saving requested changes in an ASCII text file. Each batch update performed will be saved as a separate file. Each file name will include user name or loginid and datetime stamp. This would make the search, if needed, of these files easier at the DAAC side. If requester wants to roll back or undo any changes then s/he will have to submit a new request.

5. Conclusions

Currently EOS policy states that all data products are made available to the general science community. As a consequence, the importance of mechanisms to ensure the quality of the products prior to their distribution has been recognized, and an end-to-end QA approach has been developed. This paper has described the different elements of the EOS QA approach from data production through archival that have been adopted by the AM–1 science teams and the data processing facilities. An effort has been made to archive the QA information in the metadata in a user-friendly format, to allow maximum exploitation by the science community.

EOSDIS has proven to be adaptive to evolving QA requirements (QA Metadata Update Tool). Future enhancements to EOSDIS will include granule-level data visibility and access controls, based on information contained within the QAFlags metadata, that will allow instrument team scientists and processing facility staff to temporarily hold specific data granules that require more detailed QA analysis. Another requirement recently advocated is the need for an automated method to update the QA metadata outside of the production software. This requirement was borne out of the realization that QA procedures may be automated through post-launch experience and characterization of the spaceborne instruments and of the science software used to produce the products.

QA is an evolving element within EOS. Communication is continuing among all entities (the instrument teams, the processing facilities and the developers of EOSDIS) in order to ensure that the quality of the large volume of EOS products is defined and documented. The user community will be involved in this process after launch by providing feedback from their experiences in the use of QA metadata in the data screening/ordering process and in their research. Many of the standard EOS products are new and without heritage, and may contain questionable data in the early post-launch period. Users of EOS data must now be made aware of the QA metadata associated with each product to encourage their proper utility. This information will be made available in a timely manner prior to launch within EOSDIS.

6. Acknowledgements

The authors would like to acknowledge the support of the ECS Science Office (Robert Plante) and the ESDIS Science Office (H. K. Ramapriyan), and the guidance of the QA Working Group in the development of this effort.

7. REFERENCES

[1] Asrar, G. and J. Dozier: *EOS Science Strategy for the Earth Observing System*, American Institute of Physics Press, Woodbury, N. Y., 1994.

[2] Asrar, G., and H. K. Ramapriyan: Data and Information System for Mission to Planet Earth, *Remote Sensing Reviews* 13, pp. 1-25, 1995.

[3] Gross, C.: *B.0 Implementation Earth Science Data Model, ECS Document #**420-TP-015-002***, Hughes Information Systems Company, 1997

[4] Lutz, B.: "Development of EOS Quality Assessment (QA) methodology and EOSDIS' support for QA." Proceedings of the International Society for Optical Engineering (SPIE) Meeting, San Diego, Earth Observing Systems III, 531-540, 1998

[5] Matthews, Earnest T.: *Mission Operation Procedures - Drop 4PX: A Delta Iteration, ECS Document # **611-CD-004-003***

[6] Singhal, S.: *QA Metadata Update Tool for the ECS Project*, *ECS Document **#160-WP-002-001***, Raytheon Information Systems Company, 1998.